ACOUSTIC AND LANGUAGE MODELING OF HUMAN AND NONHUMAN NOISES FOR HUMAN-TO-HUMAN SPONTANEOUS SPEECH RECOGNITION

T.Schultz and I.Rogina

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
{tanja,rogina}@ira.uka.de

ABSTRACT

In this paper several improvements of our speech-to-speech translation system JANUS on spontaneous human-to-human dialogs are presented. Common phenomena in spontaneous speech are described, followed by a classification of different types of noises. To handle the variety of spontaneous effects in human-to-human dialogs, special noise models are introduced representing both human and nonhuman noises, as well as word fragments. It will be shown that both the acoustic and the language modeling of these noises increase the recognition performance significantly. In the experiments, a clustering of the noise classes is performed and the resulting cluster variants are compared, thus allowing to determine the best tradeoff between sensitivity and trainability of the models.

1. INTRODUCTION

Recently, a large number of studies has been devoted to the task of recognizing and understanding spontaneous speech. Compared to read speech, some specific problems exist when spontaneous speech is to be recognized. The lack of fluency, one of the most important characteristics of spontaneous speech, can lead to repetitions, restarts, interjections, stuttering, hesitations, unusual stress, and wrong pronounciation. All of the above mentioned interruptions are pauses, which can be either empty (*silence*), or which can contain any kind of noise. The noise encountered in the interruptions can be classified as either nonverbal sounds produced by the human vocal tract like laughter, lip smacks, breathing, hesitations, cough, etc. so-called *human noises*, or as nonarticulatory noises, like paper rustle, key click, door slam, telephone ring, etc. so-called *nonhuman noises.* A superposition of noises and speech is not considered in this paper.

Bootstrapping our JANUS-2 speech recognizer towards spontaneous speech, the recognition performance dropped significantly compared to read speech. [1] showed that 20% of the errors between the alignment of the phonetical reference transcription and the phonetically recognized hypothesis, were due to unmodeled pause fillers and noises in the ATIS task. This suggest that the modeling of spontaneous speech events should significantly reduce the error rate. The explicit modeling of 14 human and nonhuman noises decreased the word error rate of the PHOENIX system on the Spreadsheet-Task dramatically [2]. Compared to human-to-machine tasks, e.g. ATIS, human-to-human dialogs contain a greater variety of human and nonhuman noises. Modeling these effects is extremely important for human-to-human speech recognition tasks.

2. JANUS-2 WITH A NEW DATABASE

JANUS-2 is the spontaneous speech-to-speech translation system of Carnegie Mellon and Karlsruhe University [3, 4]. It was designed as a modular system containing a speaker independent recognizer for utterances spoken in English, Spanish, and German, and a parser which analyzes the hypotheses and translates them into an Interlingua representation. German, English or Japanese text can be generated from the Interlingua representation and synthesized by a commercially available speech output device. Several algorithms are available for acoustic modeling, i.e. TDNNs, MS-TDNNs, HMM, MLP and LVQ.

JANUS-2 was extended towards spontaneous spoken human-to-human dialogs on a new database [3]. This *Appointment Scheduling* database is being collected in a similiar fashion in German, English, and Spanish. In each session, two persons are asked to schedule a fictitious meeting with their human dialog partner. The data used in the following experiments consists of 63 English dialogs. A dialog represents in the average 9 utterances. The dialogs were divided into 43 dialogs for training (387 utterances) and 20 dialogs of different speakers for testing (173 utterances). The utterances are transcribed using a set of 14 human and 23 nonhuman noise words to represent the human and nonhuman noises. In addition, the transcription format marks word fragments produced by restarts, repetitions, and interruptions as well as pauses. Including these noise words the vocabulary has a size of 865 words.

	Training	Test
dialogs	43	20
utterances	387	173
minutes of speech	62	27
words	10760	4731
noises	2383	959

Table 1: Dialog Statistics for the training and test set

Table 1 shows the properties of the task in the training and the test set.

3. ACOUSTIC MODELING

For acoustic modeling a phonetically tied SCHMM trained for speaker independent recognition was used [5]. In order to generate acoustic models for the human and the nonhuman noises, new dedicated phonemes were added to the existing set of 46 context independent phonemes. To guarantee a minimum amount of training input per model, classes of noises have to be created. Frequent human noises ("ah", breathing, lip smack, "uh", "um") and nonhuman noises (key click, paper rustle) form a class of their own. Human noises which are less frequent build the common class +human+; rare nonhuman noises are joined in the class +nonhuman+. A special class +garbage+ was introduced to handle those word fragments which were generated by restarts, repetitions, etc., and could not be modeled as regular words.

Noise	Training		Test	
	counts	%	counts	%
ah	41	0.38	22	0.47
breathing	766	7.12	294	6.21
lip smack	316	2.94	137	2.9
uh	200	1.86	80	1.69
um	195	1.81	62	1.31
+human+	67	0.62	29	0.61
key click	401	3.73	186	3.93
paper rustle	41	0.38	24	0.51
+nonhuman+	67	0.62	18	0.38
+garbage+	74	0.69	33	0.7
Silence	215	2.0	74	1.56
total	2383	22.15	959	20.27

Table 2: Frequencies of the 10 noise classes

The absolute and relative frequencies of the 10 modeled classes of noise are shown in table 2. It can be seen, that human-to-human dialogs seems to have a very high rate of noise events and this fact is reflected in the transcription of the utterances. Figure 1 illustrates the balanced occurence of the different noises in our training and test set. In contrary the transcribed utterances of the ATIS trainingset contain not nearly enough noise words to train our noise models.

4. LANGUAGE MODELING

Different types of language modeling were evaluated. In [2] noises are allowed to follow all words without language model penalties. So noise words are treated like silences. But statistics on occurrence of noise events showed, that noise some words are more probable than others at distinct locations in the utterance. Key click and breathing for example are much more common at the beginning and at the end of an utterance.

Therefore we incorporate noise models into the language model. So the noise events are modeled like regular words by applying their language model probabilities. These two types of language modeling are compared with our baseline system: a bigram language model of perplexity 44 without the noise words. The results shown in table 3, suggest that modeling the noises like regular words improves the performance moderately.

	relative reduction
like silence	6.8%
like regular words	10.9%

Table 3: Word error reduction for language modeling

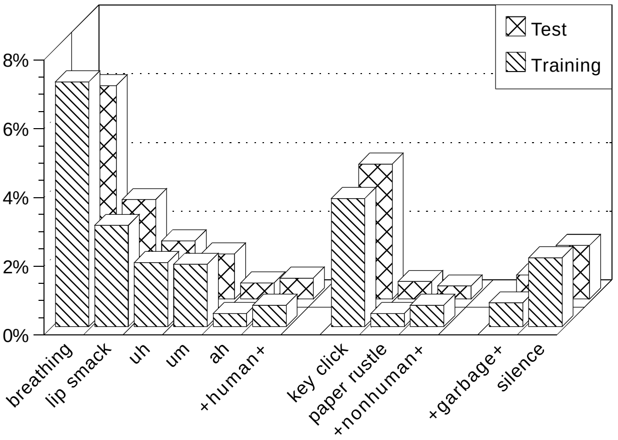

Figure 1: Percentage of noise classes in all words (monograms)

Modeling the noises like regular words requires that the noise words are distinguished from each other. So one reason for the small improvement may be a high substitution rate of noises by other noises. Table 4 shows different word error rates caused by noise models. In fact, most of the errors are due to substitutions of noises by other noises. Therefore a better modeling of the noise events, when more training data becomes available, should lead to better improvements.

word errors caused by noises	34.4%
Substitutions noises-noises	12.1
Substitutions noises-words	3.2
Substitutions words-noises	1.3
Deletions	6.17
Insertions	11.61

Table 4: Analysing the word error rate

Insertions and deletion of noises are another main source of error. But we assume that noise events do not have much semantic relevance. Therefore all noise words are eliminated from the hypotheses before parsing the recognized sentences. Because of this fact, noise-to-noise substitutions, noise insertions, and noise deletions are irrelevant in the output of the speech recognizer. Nevertheless, the minimization of deletion and insertion errors is vital to avoid continuation errors. But the main objective, if noises are stripped out from the hypotheses, is the substitution error between noise models and word models. The table shows that this kind of substitution error is only a relatively small portion of the total error. Contrary to common belief (e.g. [1]) we found, that noise models are not highly confusable with short function words.

5. CLUSTER EXPERIMENTS

5.1. Clustering the classes of noises

Although approximately 20% of all words are noises, the lack of training data remains the main problem of acoustic noise modeling. Therefore a tradeoff between trainability and sensitivity of the models had to be found for a given training set. Our experiments examined if the merging of noise models would improve the performance of the system. Therefore the 10 noise models were clustered, and different variants of the resulting clusters were compared. An agglomerative clustering algorithm was used, based on the acoustic information loss after merging two clusters of noise models. Information loss is given by the difference of entropy between the original models and the merged model, weighted by their frequencies [6]. This algorithm used a heuristic optimization, which allowed elements to be moved from one cluster to another.

Figure 2 shows the results of the clustering procedure. The cluster variants are labeled by the number of noise classes they contain. As a result, particularly rare but acousticly similar models receive more data to be trained on.

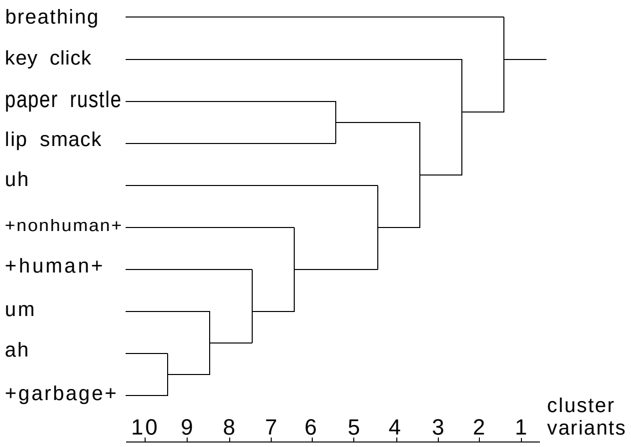

Figure 2: Result of the clustering of the 10 noise classes

5.2. Comparison of the cluster variants

For each of the resulting cluster variants 23 iterations of training were performed. Every other iteration the recognition performance was tested, using the word accuracy (WA) on the test set. Table 5 shows the averaged results of every cluster variant, using the mean WA and the best WA over all iterations.

over all Iterations	average WA	best WA
1 Cluster	45.36	47.2
2 Cluster	44.52	46.0
3 Cluster	46.34	48.7
4 Cluster	44.94	47.1
5 Cluster	46.74	49.1
6 Cluster	46.98	51.1
7 Cluster	44.52	45.2
8 Cluster	44.52	46.1
9 Cluster	46.28	49.6
10 Cluster	46.30	49.8

Table 5: Average word accuracy for all cluster variants

For the experiments the baseline system was used, so the absolute word accuracy overall is quite low. By today, the performance of the system was improved by context-dependend phonemes, data-driven codebook adaption [7], dictionary learning [8], and using morphology for language modeling [9]. JANUS-2 has at this time a word accuracy of about 66% for English and about 70% for German.

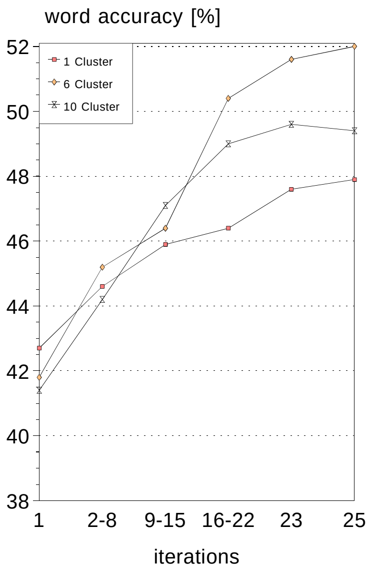

Figure 3: Word accuracy for cluster variants 1, 6, 10

Figure 3 shows the two limiting cases, and the best cluster variant. Variant 1 describes the modeling of one noise word for all human and nonhuman noises, so the maximum amount of training data was available. Variant 10 describes the opposite extreme. All noise words are trained separately, so the maximum sensitivity of the models is reached. Variant 6 yields best results for the given test set. In this variant, breathing, paper rustle, key click , lip smack and "uh" are modeled separately, the remaining noises are merged to one common cluster. This variant represents the best tradeoff between trainability and sensitivity of the noise models.

	relative reduction
1 Cluster	10 %
10 Cluster	14 %
6 Cluster	17 %

Table 6: Word error reduction for clustering

5.3. Statistical Relevance

For indicating the statistical relevance of the results we used an empirical test. For each cluster variant we used the number of misrecognized words per sentence on the test set after the 23rd iteration and performed a t-test for pairs [10] to see the significance of the differences of the mean values between the best (Cluster 6) and the other variants. Table 7 shows the mean number of misrecognized words (mean error) over all test sentences, for all cluster variants, and the significance given by the two-side probabilities p $((* : 0.01 < p \leq 0.05; ** : p \leq 0.01))$.

Cluster Variant	Iteration 23 Mean Error	Iteration 23 Significance
1 Cluster	11.52	**
2 Cluster	12.15	**
3 Cluster	11.25	**
4 Cluster	12.23	**
5 Cluster	11.06	*
6 Cluster	10.72	
7 Cluster	12.09	**
8 Cluster	11.52	**
9 Cluster	11.11	**
10 Cluster	11.00	*

Table 7: Statistical relevance of the results

6. CONCLUSION

In this paper, improvements of the JANUS-2 system towards human-to-human dialogs are presented. Analyzing the spontaneous spoken database suggests that human-to-human dialogs contain an extremely high rate of human and nonhuman noises. To model these noise events acoustic and language modeling of noises were performed. Overall, this leads to a relative word error reduction of 17 %. The lack of training data is still the main problem. When the database increase, we intend to refine the acoustic models and as consequence of better acoustic modeling the language probabilities can be applied more reliably.

7. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMFT) as a part of the VERBMOBIL project. The authors wish to thank all members of the Interactive Systems Laboratories, especially T. Kemp, P. Geutner, T. Sloboda, M. Woszczyna, and A. Waibel for useful discussions and active support.

8. REFERENCES

[1] J. Butzberger, H. Murveit, E. Shriberg, P. Price: *Modeling Spontaneous Speech Effects In Large Vocabulary Speech Recognition Applications.* SRI, Speech Research and Technology Program.

[2] W. Ward: *Modelling Non-verbal Sounds for Speech Recognition.* Proceedings of the DARPA Speech and Natural Language Workshop 1989, pp. 137-141.

[3] M. Woszczyna, N. Aoki-Waibel, F.D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, A. Waibel: *JANUS 93: Towards Spontaneous Speech Translation.* Proceedings of the ICASSP 1994, volume 1, pp 345-348.

[4] L. Osterholtz, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna: *Testing Generality in JANUS: A Multi-lingual Speech to Speech Translation System.* Proceedings of the ICASSP 1992, volume 1, pp 209-212.

[5] I. Rogina and A. Waibel: *Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems.* Proceedings of the ICASSP 1994, volume 1, pp 217-220.

[6] Kai-Fu Lee: *Context-Dependent Phonetic Hidden Markov Models for Speaker-Independent Continuous Speech.* IEEE Transactions on Acoustics, Speech, and Signal Processing (ASSP), April 1990.

[7] T. Kemp: *Data-Driven Codebook Adaption in Phonetically Tied SCHMMS.* Proceedings of the ICASSP 1995.

[8] T. Sloboda: *Dictionary Learning: Performance through Consistency.* Proceedings of the ICASSP 1995.

[9] P. Geutner: *Using Morphology Towards Better Large-Vocabulary Speech Recognition Systems.* Proceedings of the ICASSP 1995.

[10] J. Bortz: *Statistik für Sozialwissenschaftler.* Springer Berlin, Heidelberg, 1993.